Lan Wang

Nonconvex Penalized Quantile Regression: A Review of Methods, Theory and Algorithms

0.1 Introduction

Quantile regression is now a widely recognized useful alternative to the classical least squares regression. It was introduced in the seminal paper of [Koenker and Bassett, 1978b]. Given a response variable Y and a vector of covariates $\mathbf{x}$, quantile regression estimates the effects of $\mathbf{x}$ on the conditional quantile of Y. Formally, the τth $(0 < \tau < 1)$ conditional quantile of Y given $\mathbf{x}$ is defined as $Q_Y(\tau|\mathbf{x}) = \inf\{t : F_{Y|\mathbf{x}}(t) \geq \tau\}$, where $F_{Y|\mathbf{x}}$ is the conditional cumulative distribution function of Y given $\mathbf{x}$. An important special case of quantile regression is the least absolute deviation (LAD) regression ([Koenker and Bassett, 1978a]), which estimates the conditional median $Q_Y(0.5|\mathbf{x})$.

The most prominent feature of quantile regression is its ability to incorporate heterogeneity, which can arise from heteroscedastic variances or other sources beyond the commonly used location-scale models. Quantile regression allows the covariates to influence the location, dispersion and other aspects of the conditional distribution. As an example, in the analysis of the birthweight data set ([Abreveya, 2001], Chapter 1 of [Koenker, 2005]), researchers found evidence that the effects of mother's gender, race and education level on the lower tail of baby's birth weight distribution are quite different from their effects on the central part of the conditional distribution. Such heterogeneity is likely to be overlooked by least-squares regression which focuses on modeling the conditional mean. In contrast, by examining the results of quantile regression at different choices of τ, we can obtain a more accurate understanding of the effects of $\mathbf{x}$ on Y. Furthermore, in many applications, the conditional quantile is of direct scientific interest, such as the lower quantile of baby's birth weight.

Computationally, quantile regression can be formulated as a convex optimization problem where the objective function has the form of asymmetrically weighted absolute values of residuals. It can be efficiently computed via linear programming for moderately large problems, see, for example, the *Quantreg* package ([Koenker, 2008]) in R ([R Core Team, 2008]). Quantile regression enjoys several other appealing properties. It is naturally robust to outliers in the response space. The median regression is more efficient than least squares regression to estimate the conditional mean when we have symmetric heavy-tailed random errors. For any monotone function $h(\cdot)$, we have $Q_{h(Y)}(\tau|\mathbf{x}) = h\left(Q_Y(\tau|\mathbf{x})\right)$. This equivariance property usually does not hold for the conditional mean function.

The methodology and theory of quantile regression has been thoroughly studied in the classical asymptotic framework where the number of covariates p is fixed while the sample size n goes to infinity. We refer to [Koenker, 2005] for a comprehensive introduction, see also the review articles [Yu et al., 2003] and [He, 2009] for related applications. In the classical setting (fixed p), [Wang et al., 2007a], [Li and Zhu, 2008], [Zou and Yuan, 2008], [Wu and Liu, 2009], [Shows et al., 2010], [Kai et al., 2011], [Wagener et al., 2012], [Wang et al., 2013a], among others, investigated regularized quantile regression for variable selection. For unpenalized quantile regression, when p grows with n but $p = o(n)$, several authors ([Welsh, 1989], [Bai and Wu, 1994] and [He and Shao, 2000]) established useful theory for M-estimators with non-smooth objective function which applies to quantile regression; [Belloni et al., 2011] established useful asymptotic theory uniformly over quantile level τ. However, these results no longer apply when $p > n$.

Recent advances in technology has led to greater accessibility of massive data in diverse fields such as genomics, economics, finance and image analysis. In these contemporary data sets, the number of variables is often substantially larger than the sample size ($p \gg n$). For example, genomic studies often involve a small group of patients (several dozens or less) but the microarray expression levels are measured on thousands of genes. For insightful discussions on the statistical challenges of high dimensionality, we refer to the review articles of [Donoho, 2000], [Fan and Li, 2006], [Fan et al., 2014b] and [Horowitz, 2015], among others. Quantile regression enjoys two distinctive advantages for analyzing high-dimensional data.

- *Quantile-adaptive sparsity.* Most of the existing work on high-dimensional regression relies on the notion of sparisty. In the quantile regression framework, we allow both the sparsity level (the number of nonzero coefficients) and sparsity locations (which covariates are relevant) to depend on τ, the quantile level of interest. For example, the effect of gender may be relevant for modeling the lower tail of the conditional distribution of Y given $\mathbf{x}$, but not so important if we consider the conditional median. This general notion of quantile-adaptive sparsity provides a more flexible and realistic framework for modeling high-dimensional heterogeneous data.

- *Weaker regularity conditions for asymptotic theory.* For high-dimensional regression, the conditions imposed to derive the asymptotic theory are as important as the theoretical results themselves. For the theory of sparse quantile regression, we do not need to impose restrictive distributional or moment conditions on the random errors and allow their distributions to depend on the covariates. In contrast, existing theory in the literature for high-dimensional penalized least squares usually requires Gaussian or Sub-Gaussian and the vast majority requires at least second moments for the random error. We consider this as a significant advantage of sparse quantile regression as model checking is a daunting task in high dimension.

The literature on penalized quantile regression has been growing very rapidly in recent years. In this chapter, we provide a selective review of recent developments on nonconvex penalized quantile regression in the setting $p \gg n$, which requires the developments of new asymptotic theory and algorithms. It is worth emphasizing that important progress has also been made on L_1 penalized quantile regression in high dimension. [Belloni and Chernozhukov, 2011] derived nice asymptotic results uniformly in τ; [Wang, 2013] also showed that with large probability, the L_2 estimation error of L_1 penalized median regression has a near-oracle rate. Weighted L_1 penalized quantile regression was recently considered in [Bradic et al., 2011] and [Fan et al., 2014a] for efficiency or robustness consideration when $p \gg n$. We refer to the chapter by Belloni, Chernozhukov and Kato in this handbook for an in-depth review.

The rest of the chapter is organized as follows. In Section 2, we review nonconvex penalized linear quantile regression in ultra-high dimension. Section 3 discusses nonconvex penalized semiparametric quantile regression. Section 4 considers computational aspects of nonconvex penalized quantile regression. Section 5 discusses simultaneous estimation and variable selection at multiple quantiles, and two-stage analysis with quantile-adaptive marginal screening. Section 6 concludes the chapter.

0.2 High-dimensional sparse linear quantile regression.

0.2.1 Background on penalized high-dimensional regression and the choice of penalty function

We will first consider the more familiar mean regression setting and briefly discuss the general intuition behind penalization/regularization for high-dimensional regression and the main motivation for the use of nonconvex penalty functions. The same intuition and motivation apply to high dimensional quantile regression. Let $\{Y_i, \mathbf{x}_i\}$, $i = 1, \ldots, n$, be a random sample from the regression model

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta}_0 + \epsilon_i, \tag{1}$$

where $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})'$ is a vector of covariates with $x_{i0} = 1$, $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0p})'$ is a vector of unknown parameters, and the random error ϵ_i satisfies $E(\epsilon_i|\mathbf{x}_i) = 0$. In practice, it is common to standardize the design matrix such that each column (corresponding to the n observations on one covariate) has L_2 norm $\sqrt{n}$.

The primary challenge for high-dimensional regression ($p \gg n$) is that the estimation problem is ill-posed. Fortunately, in many applications it is reasonable to assume the true parameter $\boldsymbol{\beta}_0$ is sparse, that is, most of its components are zero. Hence, the regression function resides in a low-dimensional manifold. Under the sparsity assumption, the use of penalization or regularization can help achieve both estimation accuracy and interpretability. There has been a large amount of literature on penalized least-squares procedures for conditional mean regression (i.e., $E(\epsilon_i|\mathbf{x}_i) = 0$), see [Fan and Lv, 2008] for many references. The penalized least squares regression minimizes $n^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \sum_{i=1}^{p} p_{\lambda_n}(|\beta_j|)$, where $p_{\lambda_n}(\cdot)$ denotes a penalty function with a positive tuning parameter λ_n. The choice of penalty function $p_{\lambda_n}(\cdot)$ is directly related to the goal of high-dimensional regression, which are often two-fold ([Bickel, 2008]):

- *Prediction:* to provide an accurate prediction of a future observation;
- *Sparsity recovery:* to identify zero and nonzero coefficients, hence to accurately illustrate the relationship between $\mathbf{x}$ and Y.

The two objectives have subtle but important differences. In microarray studies, the first goal corresponds to constructing an effective prediction model for predicting the response of a future patient; the second goal aims to identify the set of relevant genes as therapeutic targets.

A popular choice of the penalty function is the L_1 or Lasso penalty ([Tibshirani, 1996]) for which $p_{\lambda_n}(|\beta_j|) = \lambda_n|\beta_j|$. The use of Lasso penalty achieves accurate prediction under weak conditions ([Greenshtein and Ritov, 2004]) and is computationally convenient due to the convex structure. However, it often requires stringent conditions on the design matrix to consistently identify the underlying model ([Zou, 2006]; [Zhao and Yu, 2006]). This motivates the use of nonconvex penalty, which alleviates the problem of over-penalization of L_1 penalty and can consistency identify the underlying model under much more relaxed conditions on the design matrix ([Fan and Li, 2001], fixed p). The work we review in this chapter emphasizes the second goal of sparsity recovery.

0.2.2 Nonconvex penalized high-dimensional linear quantile regression

0.2.2.1 Overview

Given a $0 < \tau < 1$, linear quantile regression imposes the model $Q_{Y_i}(\tau|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}_0(\tau)$. Equivalently, we can write the model in the form of (1) by taking $\epsilon_i = Y_i - Q_{Y_i}(\tau|\mathbf{x}_i)$, which implies that ϵ_i satisfies the quantile constraint $P(\epsilon_i \leq 0|\mathbf{x}_i) = \tau$. Note that the quantile regression coefficients are allowed to depend on τ. In the special case where the ϵ_i are independent and identically distributed, the slopes of the conditional quantiles are constant across different values of τ.

We consider estimating the τth $(0 < \tau < 1)$ conditional quantile of Y given a vector of high-dimensional covariates $\mathbf{x}$. For notation simplicity, we write $\boldsymbol{\beta}_0(\tau)$ as $\boldsymbol{\beta}_0$ when no confusion will be caused. The penalized quantile regression minimizes

$$Q(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|), \qquad (2)$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the quantile loss function (or check function), and $p_{\lambda_n}(\cdot)$ is a penalty function with a tuning parameter λ_n. The tuning parameter λ_n controls the model complexity. In the following, we simply denote λ_n by λ for notation simplicity. The role of the penalty function is to shrink the estimates of small coefficients toward zero. If the underlying model is sparse, when the penalty function and λ are appropriately chosen, many estimated coefficients will be shrunken to exactly zero which results in simultaneous estimation and variable selection.

We focus on nonconvex penalty functions here for the purpose of sparsity recovery (see Section 2.1). For theoretical development, the penalty function only needs to satisfy some general conditions (see Section 2.2.2). Two popular choices of nonconvex penalty functions are the SCAD penalty function ([Fan and Li, 2001]) and the MCP penalty function ([Zhang, 2010]). The SCAD penalty function is given by

$$\begin{aligned} p_\lambda(|\beta|) &= \lambda|\beta| I(0 \le |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \le |\beta| \le a\lambda) \\ &\quad + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \end{aligned}$$

for some $a > 2$. The MCP function has the form

$$p_\lambda(|\beta|) = \lambda\Big(|\beta| - \frac{\beta^2}{2a\lambda}\Big) I(0 \le |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \ge a\lambda),$$

for some $a > 1$. Figure 1 depicts the two penalty functions. The use of nonconvex penalty functions demands development of new asymptotic theory (see Section 2.2.3). Due to the nonsmoothness of the loss function and the nonconvexity of the penalty function, new algorithms are required to compute the penalized quantile regression estimator (see Section 4).

The problem of selecting the tuning parameter λ is crucial in practice. One popular approach is to use cross-validation to select λ. However, cross-validation is known to often lead to overfitting ([Wang et al., 2007b], [Zhang et al., 2010]). When the goal is to identify the underlying sparsity pattern, [Chen and Chen, 2008], [Wang et al., 2009a], [Kim et al., 2012] and [Wang et al., 2013b], among others, showed that a modification of BIC-type criterion achieves model selection consistency for conditional mean regression model when $p > n$. Motivated by the recent work of [Lee et al., 2013] for quantile regression, we choose the λ that minimizes

$$\text{HBIC}(\lambda) = \log\Big(\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i'\boldsymbol{\beta}_\lambda)\Big) + |\mathcal{S}_\lambda| \frac{\log(\log n)}{n} C_n, \tag{3}$$

where $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda,1}, ..., \hat{\beta}_{\lambda,p})'$ is the penalized quantile regression estimator obtained by minimizing (2) with the tuning parameter λ, $\mathcal{S}_\lambda \equiv \{j : \hat{\beta}_{\lambda,j} \neq 0, 1 \le j \le p\}$ is the estimated support of the model, $|\mathcal{S}_\lambda|$ is its cardinality, and C_n is a sequence of positive constants diverging to infinity as n increases such that $C_n = O(\log(p))$.

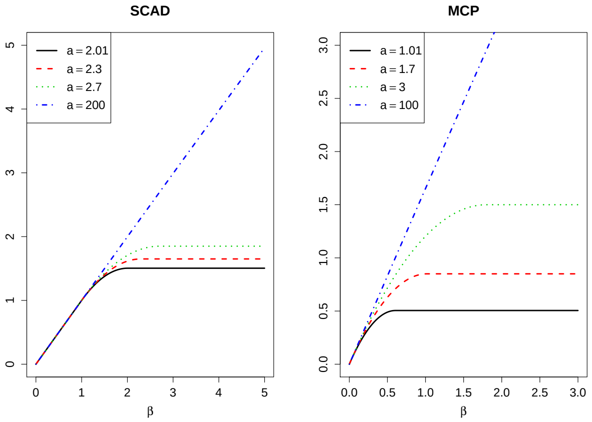

FIGURE 1
SCAD and MCP penalty functions ($\lambda = 1$)

0.2.2.2 Oracle property of the nonconvex penalized quantile regression estimator

Let $S_0 = \{1 \leq j \leq p_n : \beta_{0j} \neq 0\}$ be the index set of the nonzero coefficients in the unknown true quantile regression parameter $\boldsymbol{\beta}_0$, and $|S_0| = q_n$ be its cardinality, which may increase with n and satisfies $q_n = O(n^{c_1})$ for some $0 \leq c_1 < 1/2$. Note that both S_0 and q_n also depend on τ but we omit the dependence in the notation for simplicity. Without loss of generality, we assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', \mathbf{0}')'$, where $\mathbf{0}$ denotes a $(p_n{-}q_n)$–dimensional vector of zeros. The oracle estimator is defined as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1', \mathbf{0}')'$, where $\widehat{\boldsymbol{\beta}}_1$ is the quantile regression estimator obtained when the model is fitted using only the relevant covariates in S_0.

[Fan and Li, 2001] first studied nonconvex penalized likelihood and established the oracle property in the classical fixed p setting. An estimator of $\boldsymbol{\beta}_0$ is said to possess the *oracle property* if with probability approaching one, it estimates the zero coefficients to be exactly zero; and it asymptotically estimates $\boldsymbol{\beta}_{01}$ as efficiently as if S_0 is known in advance. [Kim et al., 2008], [Zhang, 2010], [Fan and Lv, 2011] further developed the oracle theory for nonconvex penalized least squares regression when $p \gg n$.

The techniques developed in the aforementioned literature for penalized least squares regression do not apply to quantile regression as we need to handle both the nonsmooth loss function and the nonconvex penalty function in ultra-high dimension. In particular, the penalized least squares regression problem can be written as a constrained smooth optimization problem, for which the Karush-Kuhn-Tucker (KKT) condition is sufficient ([Bertsekas, 2008]) and plays a key role in establishing the oracle theory. For quantile regression with nonconvex penalty, the KKT local optimality condition is necessary but generally not sufficient. [Wang et al., 2012] recently established the oracle theory for nonconvex penalized sparse linear quantile regression when p is allowed to grow at an exponential rate of n. They showed that the nonconvex penalized quantile regression objective function can be represented as the difference of two convex function. They then made use of a novel sufficient optimality condition for the convex differencing algorithm ([Tao and An, 1997]) and employed empirical process techniques to derive various error bounds to establish the asymptotic theory.

The penalty function $p_\lambda(t)$ is assumed to be nondecreasing and concave for $t \in [0, +\infty)$, with a continuous derivative $\dot{p}_\lambda(t)$ on $(0, +\infty)$. Assume that there exist positive constants c_2 and M such that $2c_1 < c_2 \leq 1$ and $n^{(1-c_2)/2} \min_{1 \leq j \leq q_n} |\beta_j| \geq M$. Then under some regularity conditions, for $\lambda = o(n^{-(1-c_2)/2})$, $n^{-1/2} q_n = o(\lambda)$ and $\log(p) = o(n\lambda^2)$, [Wang et al., 2012] showed that

$$P(\widehat{\boldsymbol{\beta}} \in \mathcal{B}_n(\lambda)) \to 1$$

as $n \to \infty$ where $\mathcal{B}_n(\lambda)$ is the set of local minima of the nonconvex penalized quantile objective function (2). Note that if $\lambda = n^{-\frac{1}{2}+\delta}$ for some $c_1 < \delta < \frac{c_2}{2}$, then we have $p = o(\exp(n^\delta))$, which is referred to as non-polynomial order or NP-dimensionality in the statistical literature. Also, the above oracle property is derived without imposing restrictive distributional or moment conditions on the random errors which are often required for high-dimensional penalized mean regression. More specifically, we only assume that the conditional probability density function of the random error ϵ_i, denoted by $f_i(\cdot|\mathbf{x}_i)$, is uniformly bounded away from 0 and ∞ in a neighborhood around 0 for all i. This includes, for example, the Cauchy distribution, and is much milder than the Gaussian or sub-Gaussian assumption in themean regression literature.

0.3 High-dimensional sparse semiparametric quantile regression

0.3.1 Overview

Semiparametric quantile regression is of significant importance for high-dimensional data analysis for several reasons. First of all, it possesses all the advantages we discussed earlier for the sparse quantile regression framework. Second, it incorporates nonlinear covariate effects, which often arise in real data analysis, and circumvents the curse of dimensionality associated with fully nonparametric models. Third, it alleviates the difficulty of model checking in high dimension by using a more flexible regression model.

Several authors have made important contributions to semiparametric quantile regression in the classical fixed p condition, including [He and Shi, 1996], [He et al., 2002], [Kim, 2007], [Wang et al., 2009b], [Qian and Peng, 2010], [Koenker, 2011], [Noh et al., 2012], [Zhu et al., 2010], [Bücher et al., 2014], [Yin et al., 2014], [Ma and He, 2016] among others. Recently, [Kato, 2011] studied grouped penalized quantile regression and derived a non-asymptotic bound on the estimation error; [Tang et al., 2013] considered a two-step procedure for a nonparametric varying coefficients quantile regression model with a diverging number of nonparametric functional coefficients; [Sherwood and Wang, 2016] studied partially linear additive quantile regression in the ultra-high dimension.

In the following, we focus on the high-dimensional partially linear additive quantile regression, one of the most popular seminparametric regression models. Let Y be the response variable, and let $\mathbf{x} = (x_1, ..., x_{p_n})'$ and $\mathbf{z} = (z_1, ..., z_d)'$ be p_n- and d-dimensional vectors of covariates, respectively. We assume that the τth $(0 < \tau < 1)$ conditional quantile of Y given $(\mathbf{x}', \mathbf{z}')$ is

$$Q_Y(\tau|\mathbf{x}, \mathbf{z}) = \mathbf{x}'\boldsymbol{\beta}_0 + g(\mathbf{z}), \tag{4}$$

where $g(\mathbf{z}) = g_0 + \sum_{j=1}^{d} g_j(z_j)$, with $g_0 \in \mathcal{R}$. For identification purpose, g_j is assumed to satisfy $E[g_j(z_j)] = 0$, $j = 1, \dots, d$. As an example of application in analyzing microarray data, the vector $\mathbf{x}$ may contain the gene expression levels of thousands of genes, while the vector $\mathbf{z}$ may contain clinical or environment variables of interest.

0.3.2 Nonconvex penalized partially linear additive quantile regression

Let $\{Y_i, \mathbf{x}_i, \mathbf{z}_i\}$, $i = 1, ..., n$, be a random sample generated from the partially linear additive quantile regression model in (4). We use a linear combination of B-spline basis functions to approximate $g(\cdot)$. Specifically, let $\boldsymbol{\pi}(t) = (b_1(t), ..., b_{k_n+l+1}(t))'$ be a vector of normalized B-spline basis functions of order $l + 1$ with k_n quasi-uniform internal knots on $[0, 1]$. Then $g(\cdot)$ can be approximated using a linear combination of B-spline basis functions in $\mathbf{\Pi}(\mathbf{z}_i) = (1, \boldsymbol{\pi}(z_{i1})', \dots, \boldsymbol{\pi}(z_{id})')'$. We refer to [Schumaker, 1981] for details of the B-spline construction, and the fact that there exists $\boldsymbol{\xi}_0 \in \mathcal{R}^{L_n}$, where $L_n = d(k_n + l + 1) + 1$, such that $\sup_{\mathbf{t} \in \mathcal{R}^d} |\mathbf{\Pi}(\mathbf{t})'\boldsymbol{\xi}_0 - g(\mathbf{t})| = O\left(k_n^{-r}\right)$. For ease of notation, in the sequel we use the same number of basis functions for $g_j(\cdot)$, $j = 1, \dots, d$. In practice, such restrictions are not necessary.

We are interested in the case that p_n is much larger than the sample size n. In the following, we focus on the case the number of nonlinear components d is fixed. The penalized partially linear additive quantile regression estimator minimizes

$$Q^P(\boldsymbol{\beta}, \boldsymbol{\xi}) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi}) + \sum_{j=1}^{p_n} p_\lambda(|\beta_j|), \tag{5}$$

where $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ. Denote the penalized quantile regression estimator by $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')$ and write $\hat{\boldsymbol{\xi}} = (\hat{\xi}_0, \hat{\boldsymbol{\xi}}_1', \ldots, \hat{\boldsymbol{\xi}}_d')'$ with $\hat{\xi}_0 \in \mathcal{R}$ and $\hat{\boldsymbol{\xi}}_j \in \mathcal{R}^{k_n+l+1}$, $j = 1\ldots, d$. We then estimate g_j by $\hat{g}_j(z_{ij}) = \boldsymbol{\pi}(z_{ij})'\hat{\boldsymbol{\xi}}_j - n^{-1}\sum_{i=1}^n \boldsymbol{\pi}(z_{ij})'\hat{\boldsymbol{\xi}}_j$, for $j = 1, ..., d$; and estimate g_0 by $\hat{g}_0 = \hat{\xi}_0 + n^{-1}\sum_{i=1}^n \sum_{j=1}^d \boldsymbol{\pi}(z_{ij})'\hat{\boldsymbol{\xi}}_j$. The centering of $\hat{g}_j$ is the sample analog of the identifiability condition $E[g_j(\mathbf{z}_i)] = 0$. The estimator of $g(\mathbf{z}_i)$ is $\hat{g}(\mathbf{z}_i) = \hat{g}_0 + \sum_{j=1}^d \hat{g}_j(z_{ij})$.

The practical performance of the B-spline approximation to nonlinear functions depends on the number of knots k_n. In practice, we found that a small number of knots, between 3 to 5, usually works well in a variety of settings. A high-dimensional BIC criterion, similar to that discussed in Section 2.2.1, can be used to select the tuning parameter λ. We select the λ that minimizes

$$\text{QBIC}(\lambda) = \log\left(\sum_{i=1}^n \rho_\tau\left(Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\Pi}(\mathbf{z}_i)'\hat{\boldsymbol{\xi}}_\lambda\right)\right) + \nu_\lambda \frac{\log(p_n)\log(\log(n))}{2n},$$

where ν_λ is the degrees of freedom of the fitted model, which is the number of interpolated fits for quantile regression.

0.3.3 Oracle properties

We will briefly summarize the large-sample properties of the oracle estimator and the penalized quantile regression estimator. The former is of independent interest because it allows the dimension of the linear parameter of the true underlying model to diverge with the sample size.

In model (4), it is assumed that the vector of coefficients $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}\ldots, \beta_{0p_n})'$ is sparse. Let $A = \{1 \le j \le p_n : \beta_{0j} \ne 0\}$ be the index set of nonzero coefficients and $q_n = |A|$. Without loss of generality, we assume that the first q_n components of $\boldsymbol{\beta}_0$ are nonzero and the other components are zero. As before, both A and q_n depend on τ (A may also depend on n), but we omit the dependence in notation. We write $\boldsymbol{\beta}_0 = \left(\boldsymbol{\beta}_{01}', \mathbf{0}_{p_n-q_n}'\right)'$.

The oracle estimator for $\boldsymbol{\beta}_0$ assumes the set A is known in advance, and has the form $\left(\hat{\boldsymbol{\beta}}_1', \mathbf{0}_{p_n-q_n}'\right)'$, where

$$\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\xi}}\right) = \underset{(\boldsymbol{\beta}_1, \boldsymbol{\xi})}{\text{argmin}} \frac{1}{n}\sum_{i=1}^n \rho_\tau\big(Y_i - \mathbf{x}_{A_i}'\boldsymbol{\beta}_1 - \boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi}\big), \tag{6}$$

where $\mathbf{x}_{A_1}'$,...,$\mathbf{x}_{A_n}'$ denote the row vectors of X_A, the submatrix consisting of the first q_n columns of design matrix X. Allowing q_n diverges with n such that $q_n = O\left(n^C\right)$ for some $C < \frac{1}{3}$, [Sherwood and Wang, 2016] showed that

$$\begin{aligned} ||\hat{\boldsymbol{\beta}}_{\mathbf{1}} - \boldsymbol{\beta}_{01}|| &= O_p\left(\sqrt{n^{-1}q_n}\right), \\ n^{-1}\sum_{i=1}^n \left(\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i)\right)^2 &= O_p\left(n^{-1}(q_n + k_n)\right). \end{aligned}$$

They also showed that for $l \times q_n$ matrix A_n with l fixed and $A_n A_n' \to G$, a positive definite matrix, $\sqrt{n} A_n \Sigma_n^{-1/2} \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} \right)$ converges to a multivariate normal distribution for some matrix Σ_n.

For the asymptotic theory of the non-convex penalized partially linear additive quantile regression defined by (5), [Sherwood and Wang, 2016] explored the local optimality condition of the convex differencing program ([Tao and An, 1997]; [Wang et al., 2012]) and extended it to incorporating nonparametric components. Let $\hat{\boldsymbol{\eta}} \equiv \left(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}' \right)$ be the oracle estimator. [Sherwood and Wang, 2016] showed that under some regularity condition, for either the SCAD or the MCP penalty function with tuning parameter λ, if $\lambda = o\left(n^{-(1-C_4)/2}\right)$ for some positive constant C_4, $n^{-1/2} q_n = o(\lambda)$, $n^{-1/2} k_n = o(\lambda)$ and $\log(p_n) = o(n\lambda^2)$,

$$P\left(\hat{\boldsymbol{\eta}} \in \mathcal{E}_n(\lambda)\right) \to 1 \text{ as } n \to \infty,$$

where $\mathcal{E}_n(\lambda)$ is the set of local minima of the nonconvex penalized quantile objective function in (6).

0.4 Computational aspects of nonconvex penalized quantile regression

Due to the nonsmoothness of the quantile loss function and the nonconvexity of the penalty function, computation of nonconvex penalized quantile regression estimator is significantly more challenging compared with the penalized least squares regression or unpenalized quantile regression.

Depending on the specific problem, some of the existing linear programming based algorithms may be adapted to nonconvex penalized quantile regression. We will review two such algorithms in Section 4.1. However, when p gets larger these algorithms slow down quickly. Section 4.2 reviews a new coordinate descent algorithm recently developed by [Peng and Wang, 2015], which can improves the computational speed substantially in high dimensions. We illustrate the algorithms using nonconvex penalized linear quantile regression, but they can be extended to the semiparametric quantile regression. This algorithm can also be applied to Lasso penalized quantile regression.

0.4.1 Linear programming based algorithms (moderately large p)

The first algorithm ([Sherwood and Wang, 2016]) can be implemented easily using the *Quantreg* package ([Koenker, 2008]). Hence, it has the advantage that it does not require the practitioners to do much programming on their own. For sparse regression, we often initiate the algorithm with $\hat{\boldsymbol{\beta}}^1 = 0$. For $t > 1$, let $\hat{\boldsymbol{\beta}}^{t-1} = (\hat{\beta}_1^{t-1}, \dots, \hat{\beta}_{p_n}^{t-1})'$ denote the estimator of $\boldsymbol{\beta}$ at step $t-1$. We update the estimator at step t by

$$\hat{\boldsymbol{\beta}}^t = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \Big\{ n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \sum_{j=1}^{p_n} p_\lambda' \left(|\hat{\beta}_j^{t-1}| \right) |\beta_j| \Big\}, \tag{7}$$

where if $\tilde{\beta}_j^{(t-1)} = 0$ we take the derivative $p_\lambda'(0)$ as $p_\lambda'(0+) = \lambda$. This amounts to approximating the nonconvex penalty function locally using a linear function (this step will be omitted if the penalty function itself is L_1), which is the core idea of the MM algorithm (e.g. [Lange, 2004], [Hunter and Lange, 2000], [Hunter and Li, 2005] or LLA algorithm ([Zou and Li, 2008]).

By observing that $|\beta_j| = \rho_\tau(\beta_j) + \rho_\tau(-\beta_j)$, we can formulate (7) as a weighted quantile regression problem on a set of augmented observations. More specifically, let $(Y_i^*, \mathbf{x}_i^*)$, $i = 1, \dots, (n+2p_n)$, where $(Y_i^*, \mathbf{x}_i^*) = (Y_i, \mathbf{x}_i,)$, $i = 1, \dots, n$; $(Y_i^*, \mathbf{x}_i^*) = (0, 1)$, $i = n+1, \dots, n+p_n$; and $(Y_i^*, \mathbf{x}_i^*) = (0, -1)$, $i = n+p_n+1, \dots, n+2p_n$. We solve (7) by fitting a weighted linear quantile regression model using these $n + 2p_n$ augmented observations with weights w_i^t, where $w_i^t = 1$, $i = 1, \dots, n$; $w_{n+j}^t = p_\lambda' \left(|\hat{\beta}_j^{t-1}| \right)$, $j = 1, \dots, p_n$; and $w_{n+p_n+j}^t = -p_\lambda' \left(|\hat{\beta}_j^{t-1}| \right)$, $j = 1, \dots, p_n$.

Alternatively, we can solve (7) by introducing slack variables and directly use linear programming, as the algorithm in [Wang et al., 2012]. That is, we write the optimization problem in (7) as

$$\min_{\boldsymbol{\xi}, \boldsymbol{\zeta}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (\tau \xi_i^+ + (1-\tau)\xi_i^-) + \sum_{j=1}^{p} w_j^{(t-1)} \zeta_j \Big\} \tag{8}$$

$$\begin{aligned} \text{subject to} \quad & \xi_i^+ - \xi_i^- = Y_i - \mathbf{x}_i'\boldsymbol{\beta};\ i = 1, 2, \cdots, n, \\ & \xi_i^+ \geq 0, \xi_i^- \geq 0;\ i = 1, 2, \cdots, n, \\ & \zeta_j \geq \beta_j, \zeta_j \geq -\beta_j;\ j = 1, 2, \cdots, p. \end{aligned}$$

where ξ_i^+, ξ_i^-, and ζ_j are slack variables. Note that (8) is a linear programming problem and can be solved using many existing software packages.

The above linear programming based algorithms are convenient to implement for moderately large p, but can slow down quickly when p gets larger (our own numerical experience indicates that this can happen when p is more than a few hundreds). Also, the convergence theory of these algorithms has not been investigated.

0.4.2 New iterative coordinate descent algorithm (larger p)

Motivated by the recent development of coordinate descent algorithms for penalized least squares regression (e.g., [Friedman et al., 2007], [Wu and Lange, 2008], [Breheny and Huang, 2011], [Mazumder et al., 2011], [Jiang and Huang, 2014], [Peng and Wang, 2015] recently considered iterative coordinate descent algorithm (QICD algorithm) for nonconvex penalized quantile regression, which combines the idea of the MM algorithm with that of the coordinate descent algorithm.

The QICD algorithm iterates between two steps:

- *The majorization minimization step.* The nonconvex objective function is replaced by its majorization function to create a surrogate objective function, which is updated in each iteration.
- *The coordinate descent step.* Within each iteration, the surrogate function is minimized by solving a sequence of univariate minimization subproblems, each of which minimizes along a selected coordinate with all other coordinates fixed.

The algorithm is remarkably fast as for each univariate minimizatin problem, we only need to compute a weighted median, which can be efficiently computed using quicksort or partition-exchange sort. Extending the theory of [Tseng, 2001], [Peng and Wang, 2015] established that QICD algorithm converges to a stationary point of the nonconvex penalized quantile regression objective function in (2) under some regularity conditions. The QICD algorithm is now implemented in the *QICD* package ([Peng, 2016]) and the *rqPen* package ([Sherwood and Maidman, 2016]) in R ([R Core Team, 2008]).

It is interesting to note that [Li and Arce, 2004] provided an example of using coordinate descent method for unpenalized median regression and claimed that it converges to an “inferior” solution. This example was sometimes cited as evidence against the coordinate descent algorithm. On the other hand, good empirical performance was also reported in [Wu and Lange, 2008] for a fast greedy coordinate descent algorithm for median regression. However, they have not studied the convergence theory. Our study revealed that the situation is quite favorable when the sample size is moderately large; when n is small, coordinate descent algorithms are likely to get stuck at kinks. It is helpful to examine Li and Arce's example more carefully. Li and Arce's example was based on merely 5 observations ($n = 5$). We did some calculation using their 5 data points. The global minimum is (-1.25, 0.83), which yields the objective function value 6.26; the coordinate descent algorithm gives the solution (-0.7, 1.1). Although this solution appears to be some distance away from the global minimum, it yields the objective function value 6.49. This is an example where the objective function is quite flat around the global minimum. The coordinate descent algorithm still yields a reasonable solution and may not be declared a complete failure. In fact, in their more recent paper, [Paredes and Arce, 2011] applied the coordinate descent algorithm for l_0-regularized median regression and reported positive empirical performance.

In addition to the technical arguments in [Peng and Wang, 2015], we can provide some more intuitive rationale for the good performance of the iterative coordinate algorithm with reasonably large sample size. It is motivated by the arguments in [Tseng, 2001], or more specifically his Lemma 3.1. Tseng's setup allows the penalty function part to be nonsmooth as long as it is separable; but he assumes the loss function to be smooth. When n is large, the loss functions is expected to become closer and closer to a smooth function with high probability. A consequence of this is that the directional derivative can be approximated using coordinate-wise directional derivatives with high probability (recall that in the smooth case, the vector of the derivative can be computed by the derivative with respect to each coordinate separately). Hence, his basic argument can still carry through. A longer note is available from me upon request for anyone who is interested.

0.5 Other related problems

0.5.1 Simultaneous estimation and variable selection at multiple quantiles

In some applications, researchers may be interested in simultaneous variable selection and estimation at multiple quantiles. In particular, if most of the linear covariates have zero coefficients across all the quantiles of interest, group selection is likely to help combine information across quantiles.

When $p \gg n$, [Sherwood and Wang, 2016] investigated this problem using nonconvex penalized partially linear additive quantile regression. Let $0 < \tau_1 < \tau_2 < \ldots < \tau_M < 1$ be a set of quantiles of interest, where M is a positive integer. Denote $Q_{Y_i}(\tau_m|\mathbf{x}_i, \mathbf{z}_i) = \mathbf{x}_i'\boldsymbol{\beta}_0^{(m)} + g_0^{(m)}(\mathbf{z}_i)$, where $g_0^{(m)}(\mathbf{z}_i) = g_{00}^{(m)} + \sum_{j=1}^d g_{0j}^{(m)}(z_{ij})$, with $g_{00}^{(m)} \in \mathcal{R}$ and $E(g_{0j}^{(m)}(z_{ij})) = 0$, $m = 1, \ldots, M$. Write $\boldsymbol{\beta}_0^{(m)} = (\beta_{01}^{(m)}, \beta_{02}^{(m)} \ldots, \beta_{0p_n}^{(m)})'$, $m = 1, \ldots, M$. Let $\bar{\boldsymbol{\beta}}_0^j$ be the M-vector $(\beta_{0j}^{(1)}, \ldots, \beta_{0j}^{(M)})'$, $1 \leq j \leq p_n$. The set $A = \{j : ||\bar{\boldsymbol{\beta}}_0^j||_1 \neq 0, 1 \leq j \leq p_n\}$ denotes the index set of active variables, where $||\cdot||_1$ denotes the L_1 norm.

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)'}, \ldots, \boldsymbol{\beta}^{(M)'})'$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}^{(1)'}, \ldots, \boldsymbol{\xi}^{(M)'})$. For simultaneous variable selection and estimation at multiple quantiles, we estimate $(\boldsymbol{\beta}_0^{(m)}, \boldsymbol{\xi}_0^{(m)})$, $m = 1, \ldots, M$, by minimizing

$$n^{-1} \sum_{m=1}^{M} \sum_{i=1}^{n} \rho_{\tau_m}\big(Y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(m)} - \boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi}^{(m)}\big) + \sum_{j=1}^{p_n} p_\lambda(||\bar{\boldsymbol{\beta}}^j||_1). \tag{9}$$

In the above, we use group penalty which encourages group-wise sparsity and forces the covariates that have no effect on any of the M quantiles to be excluded together, see also [Yuan and Lin, 2007], [Zou and Yuan, 2008], and [Liu and Wu, 2011]. [Sherwood and Wang, 2016] showed that the above estimation procedure enjoys the oracle property under some regularity conditions.

[Belloni and Chernozhukov, 2011] derived rates of convergence that are uniform over a continuous set of quantile indices for L_1 penalized quantile regression. [Zheng et al., 2015] studied a related but somewhat different problem. They employed adaptive L_1 penalties and proposed a uniform selector of the tuning parameter for a continuous range of quantiles levels. They derived the oracle rate of uniform convergence and weak convergence of the parameter estimators.

0.5.2 Two-stage analysis with quantile-adaptive screening

0.5.2.1 Background

In big data application, the application of penalized regression is often preceded by a screening procedure, which aims to use a computationally expedient procedure to quickly reduce the dimensionality to a moderate size, which can still be larger than the sample size n but more manageable for computation. There has been active work on variable screening for mean regression model. [Fan and Lv, 2008] proposed the sure independence screening (SIS) methodology for linear regression and established the sure screening property. See also [Fan and Song, 2009], [Hall and Miller, 2009], [Fan et al., 2011], [Bühlmann et al., 2010].

For ultra-high dimensional data, [He et al., 2013] introduced a quantile-adaptive model-free variables screening procedure, which we will briefly review below. The model-free feature is in the same spirit of that of [Zhu et al., 2011] which proposed to perform variable screening without specifying a particular model structure. This is appealing in practice due to the difficulty of validating a statistical model in high dimension. The quantile-adaptive feature is the same as what we discussed earlier for penalized quantile regression, which allows the sets of active variables to be different when modeling different conditional quantiles. It is effective for analyzing high-dimensional heterogeneous data.

0.5.2.2 Quantile-adaptive model-free nonlinear screening

At a given quantile level τ $(0 < \alpha < 1)$, we define the set of active variables

$$M_\tau = \{j : Q_Y(\tau|\mathbf{x}) \text{ functionally depends on } X_j\},$$

where $\mathbf{x} = (X_1, \ldots, X_p)'$. The variable screening procedure proposed in He, Wang and Hong (2013) ranks the importance of variables by a marginal quantile utility based on the observation

$$Y \text{ and } X_j \text{ are independent} \Leftrightarrow Q_Y(\tau|\mathbf{x}) - Q_Y(\tau) = 0, \ \forall\ \alpha \in (0,1),$$

where $Q_Y(\tau)$ is the τth unconditional quantile of Y.

We estimate the marginal condition quantile using B-spline approximation. We observe that the τth conditional quantile of Y given X_j has the expression $f_j(X_j) = \arg\min_f \mathrm{E}[\rho_\alpha(Y - f(X_j))]$. We then approximate $f_j(t)$ by $\boldsymbol{\pi}(t)'\boldsymbol{\beta}$, where $\boldsymbol{\pi}(t) = (B_1(t), \dots, B_N(t))'$ is a vector of basis functions. Let $\widehat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^N} \sum_{i=1}^n \rho_\tau(Y_i - \boldsymbol{\pi}(X_{ij})'\boldsymbol{\beta})$, and define $\widehat{f}_{nj}(t) = \boldsymbol{\pi}(t)'\widehat{\boldsymbol{\beta}}_j - F_{Y,n}^{-1}(\alpha)$ where $F_{Y,n}^{-1}(\alpha)$ is the α-th sample quantile function based on $Y_1, \dots, Y_n$. Thus $\widehat{f}_{nj}(t)$ is a nonparametric estimator of $Q_Y(\tau|\mathbf{x}) - Q_Y(\tau)$, which is expected to be close to zero if X_j is independent of Y.

The screening procedure retains all variables in the set

$$\widehat{M}_\alpha = \{1 \le j \le p : ||\widehat{f}_{nj}||_n^2 \ge \nu_n\}$$

where $||\widehat{f}_{nj}||_n^2 = n^{-1} \sum_{i=1}^n \widehat{f}_{nj}(X_{ij})^2$ and ν_n is a threshold value. A rule of thumb in practice is to rank all the features by the magnitude of $||\widehat{f}_{nj}||_n^2$ and keep the top $[n/\log(n)]$ features. [He et al., 2013] established the sure screening property of the proposed procedure, that is,

$$P\Big(M_\tau \subset \widehat{M}_\alpha\Big) \to 1, \quad \text{as } n \to \infty.$$

Hence, with probability approaching one, all important variables are retained. This is the most important property of marginal screening. Furthermore, they showed that

$$P\Big(|\widehat{M}_\tau| \le 2N^2 n^\alpha \lambda_{max}\big(\mathbf{\Sigma}\big)/\delta\Big) \to 1, \quad \text{as } n \to \infty.$$

where α and δ are some positive constants, $\mathbf{\Sigma} = \mathrm{E}\big(\mathbf{\Pi}\mathbf{\Pi}'\big)$ with $\mathbf{\Pi} = \big(\boldsymbol{\pi}(X_1), \dots, \boldsymbol{\pi}(X_p)\big)'$. If $\lambda_{max}\big(\mathbf{\Sigma}\big) = O(n^\gamma)$ for some $\gamma > 0$, then the model obtained after screening is of polynomial size with high probability. [He et al., 2013] also investigated quantile-adaptive nonlinear screening for the random censoring case.

0.6 Discussions

This chapter provides a selective overview of nonconvex penalized quantile regression in high dimension. Penalized quantile regression provides a valuable and powerful tool for analyzing high-dimensional heterogeneous data. It relies on the more flexible quantile-adaptive sparsity framework, and generally requires weaker conditions for the asymptotic theory comparing with penalized least squares regression.

Although high-dimensional data analysis has become the most active research area in statistics, there are still many challenging unsolved problems which call for the development of new methods, algorithms and theory. For example, in today's big data era, one often faces millions of observations and thousands of variables. Existing algorithms are still quite powerless with such scale of data. Developing scalable algorithms for quantile regression that can handle larger magnitudes of data is an urgent issue. Recently, [Yu et al., 2016] explored a parallel algorithm using the alternating direction method of multiplier (ADMM, [Boyd et al., 2011]) approach for large-scale non-convex penalized quantile regression and observe favorable performance when both n and p are large. Developing the methods and theory for high-dimensional quantile regression in areas such as survival analysis and longitudinal data analysis is also important. These exciting research areas pose both great challenges and opportunities.

As a referee point out, an alternative procedure to reduce the bias of Lasso is to threshold the Lasso estimator and then reestimate the model using only covariates with nonzero coefficients in the previous step. If the threshold parameter is selected appropriately, the resulted estimator also enjoys the oracle property asymptotically under regularity conditions. To achieve this, it usually requires to properly choose two regularization parameters: one for the lasso and the other one for thresholding the lasso. Thus, the tuning parameters selection is critical in such a procedure. If a covariate is mistakenly deleted in an earlier step, it will be excluded from the final fitted model. The numerical results for the mean regression model in [Wang et al., 2013c] suggested that the refitted least squares estimator based on thresholding Lasso performs similarly as the nonconvex penalized estimator in the large sample setting; although Fan and Li (2001) demonstrated in their Figure 5(c) that the hard-thresholding rule typically inflates the L_2 risk due to its discontinuity when the sample size is small.

As another referee pointed out, it was observed in [Leeb and Pötscher, 2006] and [Leeb and Pötscher, 2008] that for a general class of sparse estimators, it is impossible to consistently estimate the distribution of these estimators uniformly with respect to the unknown true regression parameter $\boldsymbol{\beta}_0$ in a small neighborhood of zero. In particular, the minimax risk behavior of this class of estimators may be undesirable if the true regression parameter $\boldsymbol{\beta}_0$ have components not exactly zero but very close to zero. This class of estimators include SCAD, Lasso, other popular shrinkage estimators such as hard-thresholded estimators, and post-model-selection estimators such as refitted regression estimator after BIC or Lasso model selection. This observation appears to have two immediate consequences on the theory and practice of penalized regression. The first is that a $\boldsymbol{\beta}$-min condition is often imposed as part of the regularity conditions for both nonconvex-penalized regression and refitted least squares regression after model selection to effectively distinguish between zero and nonzero coefficients. This conditions requires the smallest signal to decay to zero at a rate with a certain lower bound, see [Belloni and Chernozhukov, 2009], [Wang et al., 2013c] and [Fan et al., 2015], among others. The second implication is that it makes statistical inference challenging. Fortunately, the situation is not as dismal is it first looks. [Andrews and Guggenberger, 2009] noted that the existence of a uniform consistent estimator of the sampling distribution is not necessary to achieve the goal of producing a uniformly valid confidence interval. In a moving-parameter framework in which the underlying distribution is allowed to depend on n, [Hall et al., 2009] investigated using m-out-of-n bootstrap for Lasso. [Chatterjee and Lahiri, 2011] and [Chatterjee and Lahiri, 2013] showed that bootstrap and its variants can produce valid confidence intervals for a coefficient in an underlying sparse model no matter it is zero or nonzero. Alternatively, [Zhang and Zhang, 2014] and [Van de Geer et al., 2014] showed that a desparsifying approach can be used to construct asymptotically valid confidence intervals. [Belloni et al., 2015] and [Zhao et al., 2014] further derived uniformly valid inference for sparse high-dimensional quantile regression.

Bibliography

[Abreveya, 2001] Abreveya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, 25:247–257.

[Andrews and Guggenberger, 2009] Andrews, D. W. and Guggenberger, P. (2009). Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators. *Journal of Econometrics*, 152(1):19–27.

[Bai and Wu, 1994] Bai, Z. and Wu, Y. (1994). Limiting behavior of m-estimators of regression coefficients in high dimensional linear models, i. scale-dependent case. *Journal of Multivariate Analysis*, 51:211–239.

[Belloni and Chernozhukov, 2009] Belloni, A. and Chernozhukov, V. (2009). Least squares after model selection in high-dimensional sparse models. Available from https://arxiv.org/abs/1001.0188.

[Belloni and Chernozhukov, 2011] Belloni, A. and Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39:82–130.

[Belloni et al., 2011] Belloni, A., Chernozhukov, V., and Fernández-Val, I. (2011). Conditional quantile processes based on series or many regressors. Available from https://arxiv.org/abs/1105.6154.

[Belloni et al., 2015] Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102:77–94.

[Bertsekas, 2008] Bertsekas, D. P. (2008). *Nonlinear programming (third edition).* Athena Scientific, Belmont, Massachusetts.

[Bickel, 2008] Bickel, P. (2008). Discussion on the paper sure independence screening for ultrahigh dimensional feature space by fan and lv. *Journal of the Royal Statistical Society, Series B.*, 70:883–884.

[Boyd et al., 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

[Bradic et al., 2011] Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society, Series B.*, 73:325?349.

[Breheny and Huang, 2011] Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5:232–253.

[Bücher et al., 2014] Bücher, A., El Ghouch, A., Kalisch, M., and Van Keilegom, I. (2014). Single-index quantile regression models for censored data. *ISBA Discussion Paper*.

[Bühlmann et al., 2010] Bühlmann, P., Kalisch, M., and Maathuis, M. H. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. *Biometrika*, 97:261–278.

[Chatterjee and Lahiri, 2013] Chatterjee, A. and Lahiri, S. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41:1232–1259.

[Chatterjee and Lahiri, 2011] Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625.

[Chen and Chen, 2008] Chen, J. and Chen, Z. (2008). Extended bayesian information criterion for model selection with large model space. *Biometrika*, 95:759–771.

[Donoho, 2000] Donoho, D. L. (2000). High-dimensional data: the curse and blessings of dimensionality. *American Mathematical Society Conference Mathematical Challenges of 21st Century*.

[Fan et al., 2014a] Fan, J., Fan, Y., and Barut, E. (2014a). Adaptive robust variable selection. *The Annals of Statistics*, 42:324–351.

[Fan et al., 2011] Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, 106:544–557.

[Fan et al., 2014b] Fan, J., Han, F., and Liu, H. (2014b). Challenges of big data analysis. *National Science Review*, 1:293–314.

[Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle property. *Journal of the American Statistical Association*, 96:1348–1360.

[Fan and Li, 2006] Fan, J. and Li, R. (2006). Statistical challenges with high-dimensionality: feature selection in knowledge discovery. In Sanz-Solé, M., Soria, J., Varona, J., and Verdera, J., editors, *Proceedings of International Congress of Mathematicians (ICM)*, volume II, pages 595–622. European Mathematical Society, Zürich.

[Fan et al., 2015] Fan, J., Liu, H., Sun, Q., and Zhang, T. (2015). TAC for Sparse Learning: Simultaneous Control of Algorithmic Complexity and Statistical Error. *ArXiv e-prints*.

[Fan and Lv, 2008] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B.*, 70:849–911.

[Fan and Lv, 2011] Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57:5467–5484.

[Fan and Song, 2009] Fan, J. and Song, R. (2009). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38:3567–3604.

[Friedman et al., 2007] Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1:302–332.

[Greenshtein and Ritov, 2004] Greenshtein, E. and Ritov, Y. A. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988.

[Hall et al., 2009] Hall, P., Lee, E. R., and Park, B. U. (2009). Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Statistica Sinica*, pages 449–471.

[Hall and Miller, 2009] Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18:533–550.

[He, 2009] He, X. (2009). Modeling and inference by quantile regression. *Technical report*.

[He and Shao, 2000] He, X. and Shao, Q. M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73:120–135.

[He and Shi, 1996] He, X. and Shi, P. (1996). Bivariate tensor-product b-splines in a partly linear model. *Journal of Multivariate Analysis*, 58:162–181.

[He et al., 2013] He, X., Wang, L., and Hong, H. (2013). Quantile-adaptive model-free nonlinear feature screening for high-dimensional heterogeneous data. *Annals of Statistics*, 41:342–369.

[He et al., 2002] He, X., Zhu, Z., and Fung, W. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89:579–590.

[Horowitz, 2015] Horowitz, J. L. (2015). Variable selection and estimation in high-dimensional models. *Canadian Journal of Economics*, 48:389–407.

[Hunter and Lange, 2000] Hunter, D. R. and Lange, K. (2000). Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9:60–77.

[Hunter and Li, 2005] Hunter, D. R. and Li, R. (2005). Variable selection using mm algorithms. *Annals of Statistics*, 33:1617–1642.

[Jiang and Huang, 2014] Jiang, D. and Huang, J. (2014). Majorization minimization by coordinate descent for concave penalized generalized linear models. *Statistics and Computing*, 24:871–883.

[Kai et al., 2011] Kai, B., Li, R., and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of Statistics*, 39:305–332.

[Kato, 2011] Kato, K. (2011). Group lasso for high dimensional sparse quantile regression models. Available from: https://arxiv.org/abs/1103.1458.

[Kim, 2007] Kim, M.-O. (2007). Quantile regression with varying coefficients. *The Annals of Statistics*, 35:92–108.

[Kim et al., 2008] Kim, Y., Choi, H., and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103:1665–1673.

[Kim et al., 2012] Kim, Y., Kwon, S., and Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13:1037–1057.

[Koenker, 2005] Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, New York.

[Koenker, 2008] Koenker, R. (2008). quantreg: Quantile regression. r package version 5.11. *URL http://CRAN.R-project.org/package=quantre.*

[Koenker, 2011] Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian J. of Statistics*, 25:239–262.

[Koenker and Bassett, 1978a] Koenker, R. and Bassett, G. (1978a). The asymptotic distribution of the least absolute error estimator. *Journal of the American Statistical Association*, 7:618–622.

[Koenker and Bassett, 1978b] Koenker, R. and Bassett, G. (1978b). Regression quantiles. *Econometrica*, 46:33–50.

[Lange, 2004] Lange, K. (2004). *Optimization.* Springer, New York, USA.

[Lee et al., 2013] Lee, E. R., Noh, H., and Park, B. U. (2013). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109:216–229.

[Leeb and Pötscher, 2006] Leeb, H. and Pötscher, B. M. (2006). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, 22(01):69–97.

[Leeb and Pötscher, 2008] Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of hodges? estimator. *Journal of Econometrics*, 142(1):201–211.

[Li and Arce, 2004] Li, Y. and Arce, G. R. (2004). A maximum likelihood approach to least absolute deviation regression. *EURASIP Journal on Applied Signal Processing*, 12:1762–1769.

[Li and Zhu, 2008] Li, Y. J. and Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17:163 – 185.

[Liu and Wu, 2011] Liu, Y. and Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of Nonparametric Statistics*, 23:415–437.

[Ma and He, 2016] Ma, S. and He, X. (2016). Inference for single-index quantile regression models with profile optimization. *The Annals of Statistics.*

[Mazumder et al., 2011] Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: coordinate descent with non-convex penalties. *Journal of American Statistical Association*, 106:1125–1138.

[Noh et al., 2012] Noh, H., Chung, K., and Van Keilegom, I. (2012). Variable selection of varying coefficient models in quantile regression. *Electronic Journal of Statistics*, 6:1220–1238.

[Paredes and Arce, 2011] Paredes, J. L. and Arce, G. R. (2011). Compressive sensing signal reconstruction by weighted median regression estimates. *IEEE Transactions on Signal Processing*, 59:2585–2601.

[Peng, 2016] Peng, B. (2016). *QICD: Estimate the Coefficients for Non-Convex Penalized Quantile Regression Model by using QICD Algorithm.* R package version 1.0.1.

[Peng and Wang, 2015] Peng, B. and Wang, L. (2015). An iterative coordinate-descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 24:676–694.

[Qian and Peng, 2010] Qian, J. and Peng, L. (2010). Censored quantile regression with partially functional effects. *Biometrika*, 97:839–850.

[R Core Team, 2008] R Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Schumaker, 1981] Schumaker, L. (1981). *Spline Functions: Basic Theory*. Wiley: New York.

[Sherwood and Maidman, 2016] Sherwood, B. and Maidman, A. (2016). *rqPen: Penalized Quantile Regression*. R package version 1.4.

[Sherwood and Wang, 2016] Sherwood, B. and Wang, L. (2016). Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, 44:288–317.

[Shows et al., 2010] Shows, J. H., Lu, W., and Zhang, H. H. (2010). Sparse estimation and inference for censored median regression. *Journal of statistical planning and inference*, 140:1903–1917.

[Tang et al., 2013] Tang, Y. L., Song, X. Y., Wang, H. X., and Zhu, Z. Y. (2013). Variable selection in high-dimensional quantile varying coefficient models. *Journal of Multivariate Analysis*, 122:115–132.

[Tao and An, 1997] Tao, P. D. and An, L. (1997). Convex analysis approach to d.c. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22:289–355.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

[Tseng, 2001] Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475–494.

[Van de Geer et al., 2014] Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

[Wagener et al., 2012] Wagener, J., Volgushev, S., and Dette, H. (2012). The quantile process under random censoring. *Mathematical Methods of Statistics*, 21:127–141.

[Wang et al., 2009a] Wang, H., Li, B., and Leng, C. (2009a). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B*, 71:671–683.

[Wang et al., 2007a] Wang, H., Li, G., and Jiang, G. (2007a). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25:347–355.

[Wang et al., 2007b] Wang, H., Li, R., and Tsai, C. L. (2007b). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94:553–568.

[Wang et al., 2013a] Wang, H., Zhou, J., and Li, Y. (2013a). Variable selection for censored quantile regression. *Statistica Sinica*, 23:145?167.

[Wang et al., 2009b] Wang, H. X., Zhu, Z., and Zhou, J. (2009b). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, 37:3841–3866.

[Wang, 2013] Wang, L. (2013). L1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135?151.

[Wang et al., 2013b] Wang, L., Kim, Y., and Li, R. (2013b). Calibrating nonconvex penalized regression in ultra-high dimension. *The Annals of Statistics*, 41:2505–2536.

[Wang et al., 2013c] Wang, L., Kim, Y., and Li, R. (2013c). Calibrating nonconvex penalized regression in ultra-high dimension. *Annals of statistics*, 41(5):2505–2536.

[Wang et al., 2012] Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of American Statistical Association*, 107:214 – 222.

[Welsh, 1989] Welsh, A. (1989). On m-processes and m-estimation. *Annals of Statistics*, 17:337–361.

[Wu and Lange, 2008] Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244.

[Wu and Liu, 2009] Wu, Y. C. and Liu, Y. F. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19:801–817.

[Yin et al., 2014] Yin, G. S., Zeng, D. L., and Li, H. (2014). Censored quantile regression with varying coefficients. *Statistica Sinica*, 24:855–870.

[Yu et al., 2003] Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D*, 52:331–350.

[Yu et al., 2016] Yu, L., Lin, N., and Wang, L. (2016). A parallel algorithm for large-scale nonconvex penalized quantile regression. Technical report, Washington University in St. Louis and University of Minnesota.

[Yuan and Lin, 2007] Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67.

[Zhang, 2010] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942.

[Zhang and Zhang, 2014] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

[Zhang et al., 2010] Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of American Statistical Association*, 105:312–323.

[Zhao and Yu, 2006] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

[Zhao et al., 2014] Zhao, T., Kolar, M., and Liu, H. (2014). A general framework for robust testing and confidence regions in high-dimensional quantile regression. *arXiv preprint arXiv:1412.8724*.

[Zheng et al., 2015] Zheng, Q., Peng, L., and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Annals of Statistics*.

[Zhu et al., 2010] Zhu, L., Huang, M., and Li, R. (2010). Semiparametric quantile regression with high dimensional covariates. *Statistica Sinica*, 22:1379–1401.

[Zhu et al., 2011] Zhu, L. P., Li, L. X., Li, R., and Zhu, L. X. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of American Statistical Association*, 106:1464 – 1475.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

[Zou and Li, 2008] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36:1509–1566.

[Zou and Yuan, 2008] Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, 36:1108–1126.